Relationship between Molecular Connectivity and Carcinogenic Activity: A Confirmation with a New Software Program Based on Graph Theory

Davide Malacarne,[1] Raffaele Pesenti,[2] Massimo Paolucci,[2] and Silvio Parodi[3]

[1]Istituto Nazionale per la Ricerca sul Cancro, Genova, Italy; [2]Dipartimento di Informatica Sistemistica e Telematica, Facoltà di Ingegneria, Università di Genova, Genova, Italy; [3]Istituto di Oncologia Clinica e Sperimentale, Facoltà di Medicina e Chirurgia, Università di Genova, Genova, Italy

For a database of 826 chemicals tested for carcinogenicity, we fragmented the structural formula of the chemicals into all possible contiguous-atom fragments with size between two and eight (nonhydrogen) atoms. The fragmentation was obtained using a new software program based on graph theory. We used 80% of the chemicals as a training set and 20% as a test set. The two sets were obtained by random sorting. From the training sets, an average (8 computer runs with independently sorted chemicals) of 315 different fragments were significantly ($p<0.125$) associated with carcinogenicity or lack thereof. Even using this relatively low level of statistical significance, 23% of the molecules of the test sets lacked significant fragments. For 77% of the molecules of the test sets, we used the presence of significant fragments to predict carcinogenicity. The average level of accuracy of the predictions in the test sets was 67.5%. Chemicals containing only positive fragments were predicted with an accuracy of 78.7%. The level of accuracy was around 60% for chemicals characterized by contradictory fragments or only negative fragments. In a parallel manner, we performed eight paired runs in which carcinogenicity was attributed randomly to the molecules of the training sets. The fragments generated by these pseudo-training sets were devoid of any predictivity in the corresponding test sets. Using an independent software program, we confirmed (for the complex biological endpoint of carcinogenicity) the validity of a structure–activity relationship approach of the type proposed by Klopman and Rosenkranz with their CASE program. *Key words:* carcinogenicity prediction, computer-aided programs, molecular connectivity, molecular fragments, structure–activity relationships. *Environ Health Perspect* 101:332–342 (1993)

In the field of structure–activity relationship (SAR) studies, the software programs CASE (computer-automated structure evaluation) and MULTICASE, created by Klopman and Rosenkranz (*1*), represent an original approach for elucidating mechanisms of interaction between biological systems and exogenous compounds to predict the biological activities of chemicals. The strategy adopted is based on the hypothesis that molecular connectivity identifies the tridimensional structure: fragments of connected atoms and their interatomic bonds determine to a significant extent angles between pairs of contiguous atoms and their interatomic distance. The program should be able to detect, with the help of a statistical procedure, the submolecular structures that could interact with biological sites (i.e., receptors) involved in the biological process analyzed. The structure can be responsible for the biological activity of the compound (biophore) or its inhibition (biophobe). This view partially agrees with the work of Ashby and Paton (*2*), who singled out specific molecular fragments associated with genotoxicity.

The analytical capabilities of CASE increase with the amount of data input. CASE minimizes the possibility of bias due to human factors because it identifies parameters objectively, independent of human judgment. The only human operations are the choice of the data to be submitted to analysis and the interpretation of data in output. The selection of the descriptors (molecular fragments) that are used to predict biological activity is completely automated. The choice of descriptors is based on statistically significant prevalence in active or inactive molecules.

Since 1984, many studies have been published by Klopman and Rosenkranz (*3–11*) on this subject: sets of congeneric and noncongeneric compounds have been tested for several biological endpoints (mutagenicity, carcinogenicity, etc.). We have selected for discussion in this report some papers among the most pertinent to our work. Concerning predictivity, the results obtained by Klopman and Rosenkranz change for different endpoints and for different chemical classes analyzed and overall show a high level of accuracy; often, however, predictivity has been tested only in the training set or in arbitrarily built test sets.

The general strategy of CASE is known, but the detailed structure of the software is not available because it is protected by copyright. Up to now, all reports on predictivity using CASE have been published solely by the program creators or by authors using the CASE program by license or permission. Due to these restrictions, we saw the need to develop a new, completely independent program to confirm (or disprove) the validity of the type of SAR approach used by CASE.

Our software uses graph theory to reproduce basic operations characterizing the CASE program. The program associates a graph with a molecule to represent its topological properties. The program searches for subgraphs (molecular fragments) characteristic of groups of carcinogenic or noncarcinogenic compounds. To test the performance of the software, we chose the induction of tumors in rodents as a biological endpoint. Tumors are the endpoint of carcinogenesis, a complex multistage event, in which genetic alterations are only one part of the story. We used the Carcinogenic Potency Database (CPDB) (*12–15*) and the National Toxicology Program (NTP) (*16–18*) data to obtain information on rodent carcinogenicity. We divided the data into two subsets: a randomly selected learning set including 80% of the chemicals, and a nonoverlapping test set including 20% of the chemicals. An additional control analysis tested an artificially paired set of data where carcinogenicity is attributed randomly to the molecules of the training set but not to the molecules of the test set.

Methods

Software Features

To analyze the possible relationships between the structure of molecular fragments and carcinogenicity, our software analyzes the topological properties of molecular fragments using graph theory. For a detailed introduction to graph theory, see Christofides (*19*).

Graph theory is used to relate the topological properties of molecules to their possible carcinogenicity. A graph is a pair (V, E), where V is the set $\{v_i, i = 1,...n\}$ of vertices, and E is the set $\{e_{ij} = (v_i, v_j), v_i, v_j \in V\}$ of edges that express existing relations between vertices; both vertices and edges may be labeled (i.e., they may have an associated name or value). Any compound can be represented as a graph by associating the atoms with the vertices and the bonds with the edges. This kind of representation is frequently adopted in literature because it allows easy handling of the topological properties of compounds. In fact,

Address correspondence to S. Parodi, Istituto di Oncologia Clinica e Sperimentale, Viale Benedetto XV, 10 16132 Genova, Italy.

We are grateful to R. Benigni and A. Mugnoli for their useful suggestions. We thank G. Frigerio and T. Wiley for their careful assistance in preparing the manuscript. This work was supported by grant "Finalizzato CNR/ACRO," no. 92.02343.PF39(*), grant "STEP" of the European Community, no. Ct.91-0146(DTEE), and grants MURST 40–60% to S. Parodi.

Received 9 March 1993; accepted 17 June 1993.

graph theory has many applications, such as in nomenclature, coding and information processing, storage, and retrieval (*20*).

Our software system uses a fragmentation approach to determine whether subfamilies of compounds with carcinogenic activity, or lack thereof, are characterized by the presence of some common structural features (molecular fragments). A similar approach has already been applied in earlier computer-aided methods (*21–23*) for predicting different biological activities (antiarthritic–immunoregulatory effects and antineoplastic effects). In these earlier works, not all the possible fragments within a given range of nonhydrogen atoms were generated, but only a limited subset of fragments, such as augmented atoms, heteropaths, and ring fragments. A definition of these substructural units is given by Chu et al. (*22*). Our work is mainly based on the works of Rosenkranz and Klopman (*3,4*) and on the studies of Ashby (*24,25*), who has defined indicators that can be thought of as subgraphs usually present in genotoxic compounds (genotoxicity is an important component of carcinogenicity).

Essentially, the system searches all the fragments (i.e., subgraphs) of the compounds present in the training set whose activity is known, in an attempt to determine a reliable set of fragments whose presence in compounds of unknown carcinogenicity (test set) may be an indicator of their activity. In particular, the main procedure of the program that executes the fragmentation works as follows: all the fragments within a given size of each compound of the training set are produced; a unique code is associated with any fragment yielded, and, if this code is not already present in a fragment dictionary, it is inserted in the dictionary. A list of the compounds to which the fragment belongs is linked to the fragment code and it is initially filled with the code of the compound currently examined. Otherwise, if the fragment code is already present in the dictionary, only the corresponding compound list is updated. Once all the compounds of the training set have been fragmented, the system scans the dictionary by searching for the fragments that satisfy the statistical conditions (described in later).

The program was developed in standard C language, and it can be compiled on both MS-DOS and Unix architecture. The version used for the experiments described here can run on any machine with a 3.0 or later version of MS-DOS operating system, and it requires at least 4 MB of memory and 100 MB of hard disk. A typical experiment (a single run of a standard training set of 661 molecules) takes about 4 hr of computation time on a 486 machine to develop the database of significant fragments. Two additional hours are required for the statistical analysis that selects the significant fragments. The amount of time needed to determine if a new compound of a test set contains one or more of such fragments depends mainly on the compound structure; for example, the analysis of a 40-atom (nonhydrogen) compound, normally connected, takes about 5 min, whereas a 10-atom (nonhydrogen) compound takes no more than 30 sec.

The program accepts as input an ASCII file describing the structure of the compounds that will be analyzed by a connectivity matrix. A separate interface program has been developed to graphically input such structures, storing them in that ASCII file. In general, the analysis system yields synoptic report files, but it also stores information in ASCII files in which data are organized in tables; in this way such information can be easily accessed by the most database software.

Statistical Methods

After the software has considered all the molecular subunits with size between two and eight "heavy" atoms, a statistical analysis is performed to select only significant fragments. The first selection is based on the distribution of the fragments between positive and nonpositive molecules. The training set initially generates a global number of about 278,000 fragments. Of these, about 103,000 are different fragments. For the successive stages of the analysis, the software keeps only those fragments that have a probability of random association with carcinogenicity (or lack thereof) lower than 0.125 (one tailed) according to binomial distribution. We computed our statistical estimate for the tail in the direction of biological prevalence; however, statistical fluctuations can make a fragment significant in both directions (carcinogenicity or lack thereof). Therefore, conceptually, the real confidence limits have to be considered two tailed and about twice the one-tailed level of confidence. We have calculated the probability for the entire tail of the distribution to estimate statistical significance. For each monomial of the distribution we have used the classical formula:

$$\Pr(X) = \left[\frac{N!}{X!(N-X)!}\right]\left(p^X\right)\left(q^{N-X}\right)$$

where N is the number of times in which a given fragment has been generated in different molecules (trials); X is the number of times in which the fragment has been generated by positive molecules (successes); p is the probability that one fragment has been generated by a positive molecule (probability of success); its value is determined by the ratio

$$p = \frac{\text{fragments generated by positive chemicals } (\cong 159{,}000)}{\text{fragments generated by all chemicals } (\cong 278{,}000)}$$

q is the probability that the fragment has been generated by a nonpositive molecule (probability of failure = $1 - p$); and $\Pr(X)$ is the probability of X successes (single monomial).

The fragments selected in this way are labeled "activating" if their occurrence in carcinogenic chemicals is higher than the statistical limit that we established. Similarly, the fragments are labeled "inactivating" if their occurrence in nonpositive compounds is higher than the established statistical limit. In a second stage, the program removes the fragments that are redundant because they are "imbedded" in larger fragments and have identical behavior (only the subunit with smaller size is kept). At this stage the number of fragments is reduced at least 300 times in respect to the initial set of fragments generated (generally from 103,000 to 315 fragments).

A test set, a random sample of the overall data set, is tested to search each chemical for the presence of significant fragments selected in the training stage. On the basis of fragment distribution for the chemicals in the test set, a prediction of their carcinogenicity is made.

A molecule of the test set can have one or more fragments that are present in molecules of the training set. Combining the statistical significance of these fragments, we calculate an empirical index, PI (probability index), for the molecules of the test set. An example of the calculation of this simple index follows.

A molecule, X_V, of the test set contains three fragments among those ones selected as statistically significant in the training set (F_1 and F_2 "activating," F_3 "inactivating"). The fragment F_1 has been selected because it is present, in the training set, in five active molecules (A_T, B_T, C_T, D_T, E_T) and in one inactive molecule (G_T). Similarly, F_2 is contained in four active molecules (A_T, B_T, C_T, H_T), whereas the selection of fragment F_3 originates by the presence of this subunit in four inactive molecules (G_T, Q_T, S_T, T_T). The fragments F_1 and F_2 are probably related because they were generated by a similar set of molecules. To remove the redundancies, the two fragments are treated as one fragment that originates by seven chemicals (A_T, B_T, C_T, D_T, E_T, G_T, H_T). In a similar way, the information obtained from the fragments F_3 is added to create a single aggregate (A_T, B_T, C_T, D_T, E_T, G_T, H_T, Q_T, S_T, T_T), in which the ratio between molecules with

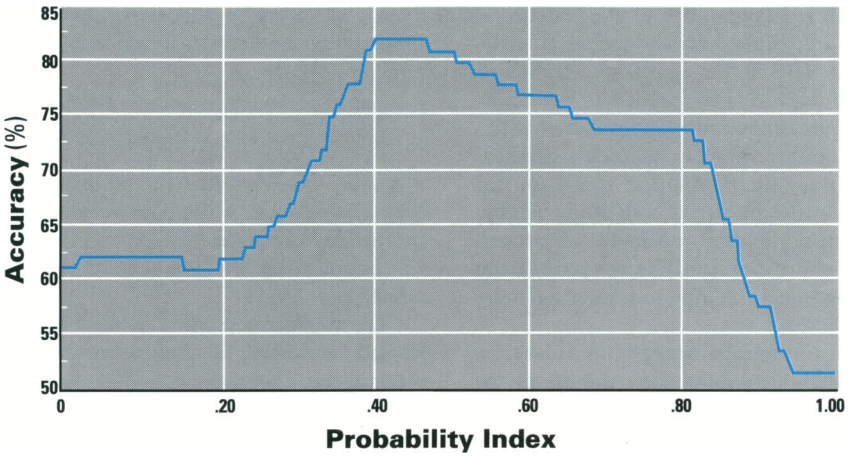


Figure 1. Behavior of the accuracy value for different probability index cutoff values in the average training set.

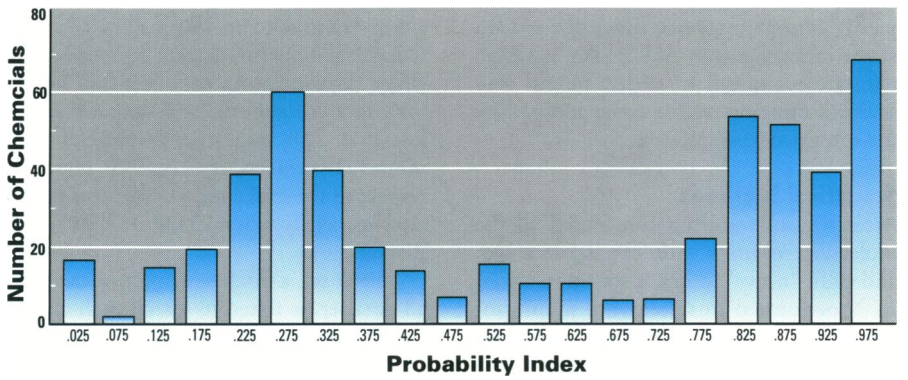


Figure 2. Distribution of probability index values for the chemicals in the average training set.

carcinogenic properties and all the molecules contributing to the evaluation is 0.6. This value is used as a PI.

A successive step is the calculation of the PI value that is used as a cut-off value to define two categories (positives and negatives) of predicted activity for the test set. This cut-off index is the value that maximizes the accuracy of the contingency table 2 × 2 (carcinogenicity or lack thereof versus predicted activity) in the training set. Accuracy in the training set as a function of the PI is illustrated in Figure 1. Levels of accuracy higher than 0.73 are obtained in the training set in a range of PI values between 0.35 and 0.8. This is because the majority of molecules have a probability index higher than 0.8 or lower than 0.35 (Fig. 2). A cut-off within this range only slightly affects the attribution to the carcinogenic or noncarcinogenic class. The average optimal cut-off value for eight runs was 0.41.

Preliminary runs of our program showed, for partial subsets of carcinogenicity data, statistical fluctuations in terms of predictivity indices. For this reason, we performed eight runs using our final database (826 compounds, 515 carcinogens and 311 noncarcinogens). For each run we randomly drew 80% of compounds for the training set and used the remaining 20% as the test set. We also performed eight paired runs using the same chemicals, but, in this case, the property of carcinogenicity in the training set was randomly attributed (pseudo-training set). The procedure for randomly selecting the chemicals for the training set and the test set imposed the condition that in both sets, 62.3% of the chemicals must be positive carcinogens. This simple procedure uses a routine of BASIC language (RANDOMIZE TIMER) as a random-number generator to assign the chemicals for the training sets and to assign the carcinogenic property in the pseudo-training sets.

To evaluate the predictivity level of our methodology, we adopted some indices that are conventionally used for diagnostic tests:

Sensitivity (SE) = [TP/(TP+FN)]100
Specificity (SP) = [TN/(TN+FP)]100
Positive predictive value (PPV) = [TP/(TP+FP)]100
Negative predictive value (NPV) = [TN/(TN+FN)]100
Observed correct predictions (OCP) = [(TP+TN)/*N*]100

where TP = true positive, FP = false positive, TN = true negative, FN = false negative, and *N* = (TP + FP + TN + FN) = number of molecules in the data set.

In addition, according to Klopman and Kolossvary (*26*), we evaluated the following two parameters:

Expected correct predictions
(ECP) = $(1 + 2 * X * Y - X - Y)100$

where X is the fraction of active molecules in the data set, and Y is the fraction of molecules predicted as active.

$$\chi^2 = N\left[\frac{TP^2}{(TP+FP)(TP+FN)} + \frac{TN^2}{(FN+TN)(FP+TN)} + \frac{FP^2}{(TP+FP)(FP+TN)} + \frac{FN^2}{(FN+TN)(TP+FN)} - 1\right]$$

Sources of Data

We gathered the carcinogenicity data analyzed here from two of the main databases: CPDB (*12–15*), in which more than 4000 experimental values are reported (1053 chemicals are considered in the database), and the NTP database (*16–18*), in which 301 chemicals have been tested with standardized protocols in mice and rats. The two databases provide qualitative and quantitative data for each experiment. We considered only qualitative results because our software can process only categorical outcomes at this time. To simplify the situation, in our first analysis we used only binary data: we classified the experimental results for each chemical as "positive" or "nonpositive." To this end, we arbitrarily fixed criteria to make a binary outcome. Table 1 shows the rules adopted for CPDB data, and Table 2 describes the rules used for NTP data. The two databases overlap extensively due to the fact that NTP data (except for most recent experiments) are already contained in CPDB. For only a few chemicals was there incomplete agreement between the two sources: Table 3 considers all the possible combinations of matched results.

A large portion of the compounds for which there are data available in the two databases is included in our database. No intentional selection was performed. We discarded 50 (4.4%) chemicals with uncertain carcinogenicity status (not classified according to Tables 1–3); 263 (23.1%) chemicals were excluded for one or more of the following reasons: 1) administered in mixture; 2) less than three "heavy" atoms; 3) molecules too large for the input interface (more than 50 heavy atoms); 4) contained unusual atoms (chemicals containing only H, C, S, N, Cl, O, Na, F, Br, P were included in the database); 5) difficulty finding the structural formula. Our program can currently analyze 826 chemicals. The CAS numbers of these chemicals are given in Appendix A.

Results

The fragmentation stage of the process produces about 278,000 fragments (average of 8 runs), adding up all the fragments produced for each molecule; of these, about 103,000 are different fragments. From the analysis of their occurrence and after removal of redundant fragments, on the average, 315 fragments significantly associated with carcinogenicity or lack thereof ($p<0.125$ according to binomial distribution) are kept for the successive steps of the analysis. The number of fragments is significantly lower for the paired training sets with a random attribution of carcinogenicity: on average, 174 fragments are selected. Detailed features of the data analyzed are summarized in Table 4. We also counted the fragments generated with a threshold of statistical significance at $p<0.01$. In this case, the training set of all the 826 chemicals in our database generated 50 fragments, whereas 6 pseudo-training sets (see Methods) of 826 chemicals generated an average of only 11.8 fragments. Examining the distribution of the fragments shown in Appendix B, we observe that the most common size is 4 heavy atoms (15 fragments), although sizes between 3 and 7 are also relatively common (5–10 fragments). Only two significant fragments of eight heavy atoms and only one fragment of two heavy atoms are present.

The 315 fragments obtained from the training stage are prevalently "inactivating" (60.6%), and only 39.4% are "activating." This fact may be due to the ratio between fragments generated from carcinogens and noncarcinogens in the database studied. In our global database we have more carcinogens (62.3%) than noncarcinogens (37.7%). However, noncarcinogens have an average size larger than carcinogens (15.1 heavy atoms versus 13.0 heavy atoms). Most likely for this reason, out of the total number of generated fragments (redundant fragments included), 57.0% come from carcinogens and 43.% from noncarcinogens. Figure 3 shows the distribution of the occurrences of 103,000 fragments of the average training set. In the case of negative fragments, those present in three noncarcinogens reach our established limit of statistical significance ($0.43^3<0.125$). This is not the case for positive fragments ($0.57^3>0.125$). For a positive fragment to become significant, it has to be present in at least four carcinogens ($0.57^4<0.125$). As shown in Figure 3, many more fragments are present at least three times than those present at least four times. Statistically significant negative fragments can be sorted from a larger set than statistically significant positive ones. As a consequence, even if we start with more positive (57%) than negative fragments (43%), we end up with 60.6% statistically significant negative fragments and 39.4% statistically significant positive ones (in the final set of 315 statistically significant different and nonredundant fragments).

Among the 315 significant and nonredundant fragments, similar (not identical), related fragments are still present, but the possible bias that they could introduce in terms of predictivity is lessened by the statistical treatment described in the previous section. These fragments generate the predictions of carcinogenicity or lack thereof for the test sets. For each run, a 2×2 contingency table is created and all the most important indices of qualitative predictivity are calculated.

Table 5 shows the contingency table obtained from the average data of eight runs for the compounds in the training sets where real experimental carcinogenicity data have been used. All the indices calculated

Table 1. Criteria used to define categories of carcinogenicity: CPDB data

Statistical significance	Authors' opinion[a]	Category[b]
$p>0.1$	–, NE	NP
	e, p, a, c, +	NC
$0.1>p>0.01$	–	NP
	e, p, a, NE	NC
	c, +	P
$p<0.01$	–	NP
	e, p	NC
	a, c, +, NE	P

[a]These notations are used in the CPDB (*13*): **a**, National Cancer Institute (NCI) or NTP evaluation is that the incidence of tumors at that site(s) was associated with administration of the compound. This code is used for technical reports before March 1986; **c**, NTP evaluation is clear evidence of carcinogenic activity. For NCI/NTP reports before March 1986, **c** indicates that the evaluation was carcinogenic; **e**, NTP evaluation is equivocal evidence of carcinogenic activity: studies that are interpreted as showing a marginal increase of neoplasms that may be chemically related; **p**, NTP evaluation is some evidence of carcinogenic activity: studies that are interpreted as showing a chemically related increased incidence of neoplasms (malignant, benign, or combined) in which the strength of the response is less than that required for clear evidence; **+**, author in general literature evaluated site as positive; **–**, in the general literature the author evaluated site as negative. NTP evaluation is no evidence of carcinogenic activity: studies that are interpreted as showing no chemically related increases in malignant or benign neoplasms; NE, no evaluation for NTP and general literature.

[b]P, positive; NC, not classified; NP, nonpositive. A chemical that could be defined as positive at least in a single species, in a single sex, in a single site, was defined as positive. A chemical that could be defined as nonpositive in all sites was defined as nonpositive. Chemicals with a mixture of not classified and nonpositive evaluations were discarded as equivocal.

Table 2. Criteria used to define categories of carcinogenicity: NTP data

Class[a]	Category[b]
A, B, C, D	P
E	NC
F	NP

[a]These notations are used in the NTP database to define the effect across the species, sexes, and tissues (*17*): A, carcinogenic in both species; B, carcinogenic in single species with two or more tissues affected; C, carcinogenic in single species with a single tissue affected; D, carcinogenic in single sex of a single species with a single tissue affected; E, equivocal study providing equivocal evidence of carcinogenicity; F, chemical associated with noncarcinogenicity.

[b]P, positive; NC, not classified; NP, nonpositive.

Table 3. Criteria used to define categories of carcinogenicity: possible overlapping

CPDB status	NTP status	Category[a]
P	P	P
P	NC	P
P	NP	P
NC	P	P
NC	NC	NC
NC	NP	NP
NP	P	P
NP	NC	NP
NP	NP	NP

[a]P, positive; NC, not classified; NP, nonpositive.

Table 4. Detailed features of the training sets (average of eight runs)

Number of compounds	661
Percentage of positive compounds	62.3
Total fragments generated	277,723
Percentage of positive fragments	57.1 (62.7)[a]
Total different fragments generated	103,490
Statistically significant fragments (after removal of redundance)	315 (174)[a]
Percentage of positive fragments	39.4 (37.5)[a]

[a]Average value of the eight paired runs with carcinogenicity randomly attributed.

Table 5. Software prediction for the training sets

	Carcinogens	Noncarcinogens	Total
Predicted positives	267.6	39.5	307.1
Predicted negatives	45.9	168.6	214.5
Total	313.5	208.1	521.6

Sensitivity	85.4%
Specificity	81.0%
Positive predictive value	87.1%
Negative predictive value	78.6%
Expected correct predictions[a]	51.8%
Observed correct predictions	83.6% (± 0.38)
χ^2	227.7 ($p < 10^{-6}$)

Carcinogenicity attributed according to Tables 1–3; average of eight runs (± SE).
[a]As defined in Klopman and Kolossvary (*26*).

Table 6. Software prediction for the training sets

	Carcinogens	Noncarcinogens	Total
Predicted positives	157.9	18.5	176.4
Predicted negatives	60.3	140.3	200.6
Total	218.2	158.8	376.9

Sensitivity	72.4%
Specificity	88.3%
Positive predictive value	89.5%
Negative predictive value	69.9%
Expected correct predictions[a]	49.5%
Observed correct predictions	79.1% (± 2.04)
χ^2	136.0 ($p < 10^{-6}$)

Carcinogenicity randomly attributed; average of eight runs (± SE).
[a]As defined in Klopman and Kolossvary (*26*).

Table 7. Software prediction for the test sets

	Carcinogens	Noncarcinogens	Total
Predicted positives	58.9	23.3	82.1
Predicted negatives	17.9	26.5	44.4
Total	76.8	49.8	126.5

Sensitivity	76.7%
Specificity	53.3%
Positive predictive value	71.7%
Negative predictive value	59.7%
Expected correct predictions	53.2%
Observed correct predictions	67.5% (± 1.32)
χ^2	11.9 ($p < 0.0006$)

Carcinogenicity attributed according to Tables 1–3; average of eight runs (± SE).
[a]As defined in Klopman and Kolossvary (*26*).

Table 8. Software prediction for the test sets

	Carcinogens	Noncarcinogens	Total
Predicted positives	30.1	21.0	51.1
Predicted negatives	29.0	20.0	49.0
Total	59.1	41.0	100.1

Sensitivity	50.9%
Specificity	48.8%
Positive predictive value	58.9%
Negative predictive value	40.8%
Expected correct predictions	50.2%
Observed correct predictions	50.1% (± 1.95)
χ^2	0.00 ($p < 1$)

Carcinogenicity randomly attributed (in the training sets); average of eight runs (± SE).
[a]As defined in Klopman and Kolossvary (*26*).

seem to show a high level of predictivity. However, even the indices obtained with the eight training sets where carcinogenicity was randomly attributed (Table 6) show a high predictivity performance. It is clear that the results obtained are not due to the predictive capability of the program but mainly to the many degrees of freedom existing in the system. These degrees of freedom allow for an *a posteriori* adaptation of the program to the pattern of positive and negative data in the training sets. In conclusion, the training sets cannot be used for an assessment of predictivity. It must be noted that the pseudo-training sets generate less "significant" fragments than the real training sets. As a consequence, there are fewer chemicals associated with a positive or negative prediction (376.9) in respect to the real training sets (521.6).

Table 7 shows the contingency table obtained for an average of eight test sets. The level of accuracy (67.5%) is significantly higher (p-0.0006) than the expected level, based on the hypothesis of no association between connectivity and carcinogenicity (53.2%). The results obtained when the training sets with carcinogenicity randomly attributed are used to predict the same test sets (Table 8) do not show any association. These results and the previous observation that for a random attribution of carcinogenicity, about 55% of apparently significant fragments are generated in respect to a real training set, strongly suggest that connectivity is associated only with a real biological property and not with a randomly distributed simulated property.

Among the 165 chemicals of the test sets: 1) 32.4% (average of eight runs) contained only statistically significant positive fragments and were predicted with an accuracy of 78.7%; 2) 24.4% of the chemicals contained only statistically significant negative fragments and were predicted with an accuracy of 60%; 3) 19.8% of the chemicals contained both statistically significant positive and negative fragments and were predicted with an accuracy of 59.3%; 4) 23.3% of the chemicals contained no statistically significant fragments (70.8% of these chemicals were carcinogens and 29.2% were noncarcinogens), thus preventing a prediction of carcinogenicity.

Of those chemicals without statistically significant fragments, the ratio between carcinogens and noncarcinogens (70.8/29.2) is higher than the ratio present in the global database (62.3/37.7). This result can be explained by the fact that among the 315 statistically significant fragments selected by the program, more negative fragments (60.6%) than positive fragments (39.4%)

are detected. For this reason, perhaps, we more often detected noncarcinogens than carcinogens. This could explain the enrichment in carcinogens among the molecules not associated with significant fragments.

Discussion

The major drawback to this type of automated analysis is the number of elementary operations performed and the quantity of memory needed. Determining the largest common subgraph between two graphs is a nonpolynomial task and requires time that exponentially depends on the size of the graphs and subgraphs involved. Fortunately, some characteristics of the chemical compounds partially simplify this otherwise formidable task: 1) the maximum number of edges converging at a node is usually small (around four); 2) the number of atoms in the compounds of our database is relatively small: the average number of heavy atoms (nonhydrogen) per compound is 13.8, and the largest compound contains 48 heavy atoms (see Fig. 4); 3) the maximum size of the searched fragments was limited to eight heavy atoms. As can be observed in Figures 5 and 6, fragments of greater size tend to appear in large numbers, but each of them tend to be present in too few compounds to be statistically significant. We have also observed that in our database, the information (associated with carcinogenicity or lack thereof) related to fragments of size 9 is redundant in respect to the information of smaller sizes in 100% of the cases (data not reported).

Finally, thus far, the adopted technique of representation of molecular fragments does not make a distinction among steric isomers; such cases will be dealt with in a future improvement to the system.

We have described the method for calculating our PI value in Methods. We used the PI value as a discriminant for deciding if a molecule of the test set will be predicted to be a carcinogen or a noncarcinogen. The strategy adopted prevents strongly related fragments from contributing to the analysis as independent fragments. In this way the informative content of a single chemical in the training set can have only one unit weight: we thus avoid the introduction of a bias of redundancy resulting from the multiplication of information related to a single molecule.

This strategy can introduce a different potential bias for a subset of molecules with different active substructures all common to the same molecules: in this case the index calculated can be underestimated. However, in our opinion, adding up the contributions of highly correlated fragments would cause more distortion than discarding multiple contributions present in the same molecule.

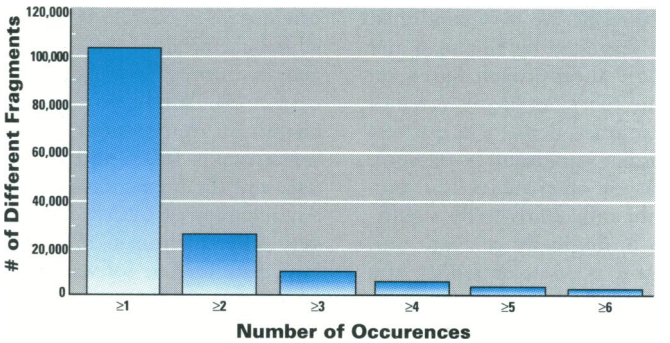


Figure 3. Fragment occurrences for 661 chemicals (average training set).

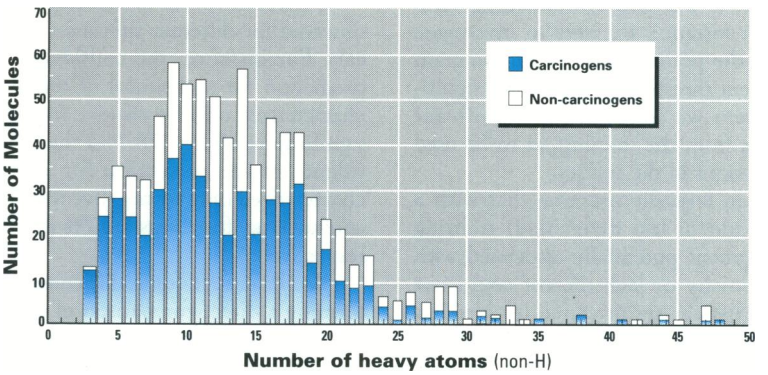


Figure 4. Size of the molecules present in the database.

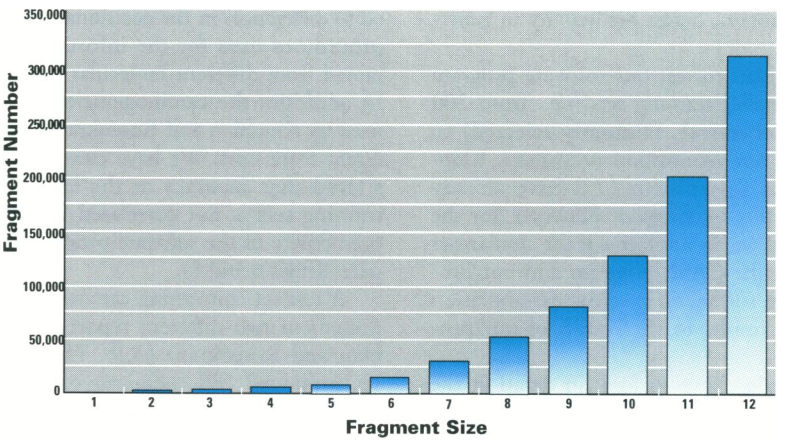


Figure 5. Global number of different fragments, according to their size. Results are for a set of 661 randomly selected chemicals.

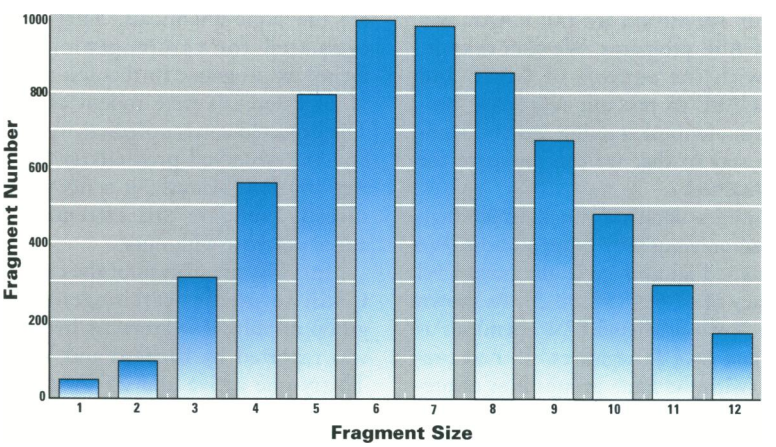


Figure 6. Number of different fragements present in at least five molecules, according to their size. Results are for a set of 661 randomly selected chemicals.

As a general result, we have confirmed what has been suggested by Klopman and Rosenkranz (*4*): an approach based on molecular connectivity can predict carcinogenicity. The results obtained in our test sets are statistically significant (p-0.0006). We believe that the observed levels of predictivity are not only statistically significant but also biologically relevant and potentially useful as one component of a spectrum of information that can contribute to hazard evaluations. Our initial work is promising, but we must test the software in additional experiments to develop it as a predictive toxicology system. For instance, we have to investigate in detail the performance of our program for different thresholds of statistical significance when we are selecting significant fragments from the training set to be used for predictions in the test set.

We can logically presume that with a smaller (and/or less diversified) training set, a fragment potentially associated with carcinogenicity or lack thereof could not reach statistical significance (or reach a more equivocal statistical significance). Therefore, we would expect that the percentage of nonassessable chemicals should decrease for a larger training set, and we should obtain better predictivity in general.

We plan to test our software program using smaller training sets (i.e., from 200 to 400 chemicals randomly selected) to verify if our assumption is correct. Klopman and Rosenkranz (*11*) have already verified this assumption. However, for the moment, we do not know if the similarities between the CASE program and our program are sufficient to allow extrapolation of their results to the results of our program.

We also have to look in detail at the fragments selected as significant to comment about their biological plausibility and compare them with the alert structures of Ashby (*2,16,17,18,24,25*) and also with fragments identified by the CASE and MULTICASE programs. We plan to coordinate with the authors of CASE and MULTICASE to test our respective programs with identical training sets and identical test sets so that we can compare the results obtained.

We used a database much larger than those used previously by other authors. We have obtained an average (eight runs) level of accuracy of 67.5% (SE, ±1.3). As shown in Table 7, we predicted 82.1 chemicals as positive and 44.4 as negatives. If these predictions (with the same proportions of predicted positives and negatives) had been based only on chance, the level of accuracy would have been 53.2% (ECP value). In our database, the prevalence of positive carcinogens is 62.3%. If we had predicted all the chemicals of the test sets as carcinogens, we would have obtained an accuracy of 62.3%. When you predict that all chemicals are potential carcinogens, the sensitivity is 100% and the specificity is 0%, and the prediction is not very useful. An accuracy of 62.3% is apparently not very different from 67.5%, but we would anticipate for our software program levels of accuracy in the range of 65–70% at a ratio of carcinogens/noncarcinogens of 50/50, or even 38/62. We plan to perform these experiments in a future study.

Different levels of predictivity were observed for different subclasses of chemicals. For instance, the confidence of the prediction for a chemical of the test sets, characterized only by positive fragments, is significantly higher (78.7%) than the confidence of the prediction for a chemical characterized only by negative fragments or contradictory fragments (60.7% and 59.3%, respectively).

We have met some difficulties in performing a direct comparison of our results with the results obtained by CASE. At the level of the training set, accuracy was higher (~95%) for CASE (*8,9*) than for our program. This difference is probably related to differences in the decisional-statistical procedures used for the information obtained from different molecular fragments. In addition, the carcinogenicity database used by Klopman and Rosenkranz was different from ours. We have clearly demonstrated that accuracy at the level of the training sets is not correlated to the real predictivity of the software program (compare Tables 6 and 8).

A test set concerning carcinogenicity is present in two different reports by Klopman and Rosenkranz (*8,9*). The training set contained 189 chemicals of the NTP study (50.2% active, 22.2% marginally active, and 27.5% noncarcinogens). The rodent carcinogens (or noncarcinogens) considered in the test sets of the two papers are the same chemicals. They had been evaluated for carcinogenicity in the GeneTox program. In this test set, 23 out of 24 chemicals were rodent carcinogens. The expected correct predictivity was 92%, and the observed predictivity (accuracy) was 100%. Obviously, it is not possible to directly compare this extremely unbalanced database with ours.

In 1990, an analysis of the capability of CASE to predict carcinogenicity for a group of polycyclic aromatic hydrocarbons was reported by Richard and Woo (*27*). Thirty-one active and 25 inactive PAHs were used in the training set ("LEARN"), and 9 active and 15 inactive PAHs were used in the test set ("VALIDATE"). The authors reported an accuracy of 75% (SE, 89%; SP, 67%). In a recent publication (*28*), results concerning the predictive capabilities of CASE were reported for a group of chemicals for which carcinogenicity data recently became available (NTP studies). Out of 25 chemicals predicted by CASE, 17 were carcinogens and 8 were noncarcinogens (6 equivocals omitted). The degree of accuracy was 64% (SE, 59%; SP, 75%). Obviously, these results are from a small test set, not directly comparable with ours.

Among the works published by Klopman and Rosenkranz, a larger database (more similar to our database) was used to predict mutagenicity in *Salmonella*. In a recent study (*1*), Klopman and Rosenkranz used mutagenicity data from the GeneTox program and NTP studies to perform the analysis. The training set was built using GeneTox mutagenicity data, and the test set was built using NTP mutagenicity data. Chemicals present in both the databases were not submitted to CASE and MULTICASE analysis. In this way, the training set contained 450 mutagens, 253 marginally active mutagens, and 123 nonmutagens, whereas the test set contained 63 mutagens, 21 marginally active mutagens, and 61 nonmutagens. The highest level of predictivity obtained using the MULTICASE program was about 80%, opposed to an expected correct prediction of about 50%. According to Ashby and Tennant (*29*), mainly electrophiles (directly or after metabolic activation) are involved in *Salmonella* mutagenicity. It is reasonable to think that mutagenicity in *Salmonella* should be more easy to predict than the complex endpoint of carcinogenicity: phenomena such as promotion, clonal expansion, remodeling, tissue necrosis and regeneration, and modulation of proliferation, apoptosis, and differentiation are clearly involved in the carcinogenic process, but not in mutagenicity in *Salmonella* or in other short-term tests of genotoxicity. We would expect a wider and more heterogeneous spectrum of molecular fragments to be involved in carcinogenicity than in genotoxicity. In the future, we will have to apply our software program not only to carcinogenicity but also to mutagenicity in *Salmonella* to test our hypothesis that it is in general easier to predict genotoxicity than carcinogenicity.

After analyzing recent studies evaluating the qualitative correlation between short-term tests for genotoxicity and carcinogenicity (*30,31*), we conclude that accuracy is in the range of 56–62%. It seems reasonable that short-term genotoxicity tests can reflect irreversible alterations in the genome during carcinogenesis. On the other hand, short-term tests should not be able to monitor nongenotoxic events (for instance, those events linked to pro-

motion and clonal expansion of preneoplastic cells). The fact that the predictivity of molecular connectivity is better than the predictivity of short-term genotoxicity tests suggests that molecular connectivity can detect not only electrophilic fragments, like the ones described by Ashby et al. (*2,16–18,24,25*), but also fragments linked to nongenotoxic effects (promotion, modulation of differentiation, etc.). An alternative explanation of this difference in accuracy could be related to the fact that nongenotoxic carcinogens may be more abundant in the databases used to assess the predictivity of short-term tests (*30,31*) than in our larger database. In the future we will investigate the predictivity of molecular connectivity for genotoxic and nongenotoxic carcinogens.

We have discussed the predictive capability of short-term genotoxicity tests. How much higher would this predictivity be with a test biologically closer to carcinogenicity in rodents? We can partially answer this question. The endpoint of carcinogenicity in a single species of small rodents is not very different in the evolutionary scale from the endpoint of carcinogenicity in at least one of two closely related species. If our endpoint is now only in mice or rats, we can predict carcinogenicity in one species with carcinogenicity in the other. For the database of Gold et al. (*12–15*), a concordance of 75% between rat and mouse studies has been reported (*32*), and for the chemicals of the NTP studies, a concordance of 74% has been reported (*33*); the predictivity of molecular connectivity is only moderately lower than the values reported above. This can be considered an additional indication of the good behavior of our parameter. We will have to confirm this impression in future experiments using only mouse data or rat data.

Within the framework of hazard evaluation, we believe that the computerized SAR approach should be given a weight similar to that of a standard short-term test in a multifactorial analysis of the carcinogenic potential of a given chemical. With regard to genotoxicity and carcinogenicity, Ashby (*34*) has pointed out that some fragments detected as significant by Klopman and Rosenkranz (and likewise by us) could not stand an in-depth analysis performed by a human expert, considering both biological and chemical specific arguments. We agree with this observation. Because we found in the pseudo-training sets a number of apparently significant fragments equal to about 55% of the statistically significant fragments found in the real training sets, we suspect that (as a first approximation) about half of the fragments defined as significant according to our statistical threshold ($p<0.125$, one tailed) are spurious. According to our analysis, only about 50% of apparently significant fragments emerging from a training set can be fragments of real biological significance. The remaining 50% is probably generated by chance and can also be present in a pseudo-training set in which carcinogenicity is assigned randomly. The level of predictivity reached in our experiments is probably due to a mixture of approximately 50% predictive fragments and approximately 50% of noise fragments. We think that fragments suggested as significant by our software program should be considered only as candidates for biological significance, but are by no means foolproof biological indicators of carcinogenicity. Their probability of being significant is higher, as expected, when we select a more severe statistical threshold. As a consequence of these considerations, a new potentially significant fragment detected by our software program is only submitted to the attention of investigators as a possible fragment characterizing a subfamily of molecules, potentially responsible for their common carcinogenic activity. Additional biological and chemical considerations could lead to the acceptance or rejection of the fragment as biologically significant. For instance, if the chemicals considered are similar procarcinogens, a similar metabolism should generate similar proximate carcinogens and perhaps also similar DNA adducts.

There are also cases in which it is impossible to reach a definite conclusion. Statistical significance is only one factor; however, when the statistical threshold is much more severe ($p<0.01$ instead of $p<0.125$), the number of significant fragments generated in a real training set is four to five times larger than the number of significant fragments generated in a pseudo-training set (against a ratio of 2/1 for the threshold, $p<0.125$). Fragments with a higher statistical significance deserve priority in subsequent biological investigations with the aim of confirming or disproving the existence of a new molecular structure relevant for carcinogenicity or genotoxicity. On the other hand, the information obtained with the threshold $p<0.125$, while less significant than the information obtained with the threshold $p<0.01$, still allowed us to make predictions about a much larger fraction of chemicals. For this reason, the threshold $p<0.125$ was selected for the general predictivity study presented here.

We have used the overall evidence of carcinogenicity in at least one species, one sex, and one tissue, without any consideration about carcinogenic potency to determine whether or not a chemical is a carcinogen (yes or no). In the future we plan to stratify our database according to spectrum of carcinogenicity (large spectrum, narrow spectrum), as suggested by Tennant (*35*) and perhaps take into consideration different ranges of potency. A subfamily of chemicals sharing a common chemical fragment could also display a relatively homogeneous behavior in respect to a different subfamily sharing a different fragment.

Finally, in conclusion, we have confirmed that with a large database, using an independent software program, SAR approaches based on the computer-automated detection of molecular fragments statistically associated with a given biological property can be used to predict carcinogenicity in rodents. We are not aware of other independent validations of this type of SAR approach.

Appendix A: CAS number of the compounds used in the analysis (826 chemicals). Chemicals are listed in numerical order.

50-06-6	60-80-0	79-46-9	96-45-7	108-60-1	132-27-4	443-48-1	628-36-4	1836-75-5	5036-03-3	16699-10-8	39801-14-4
50-07-7	61-76-7	80-08-0	97-00-7	108-78-1	132-98-9	471-29-4	628-94-4	1867-73-8	5131-60-2	16813-36-8	40548-68-3
50-18-0	61-82-5	80-33-1	97-16-5	108-88-3	133-06-2	443-72-1	630-20-6	1897-45-6	5164-11-4	17026-81-2	40580-89-0
50-23-7	61-94-9	80-62-6	97-18-7	108-95-2	133-07-3	446-86-6	632-99-5	1912-24-9	5208-87-7	17157-48-1	42011-48-3
50-24-8	62-44-2	81-07-2	97-56-3	109-69-3	133-90-4	470-82-6	634-93-5	1934-21-0	5307-14-2	17608-59-2	42579-28-2
50-29-3	62-53-3	81-16-3	97-59-6	109-84-2	134-29-2	474-25-9	636-21-5	1936-15-8	5461-85-8	17673-25-5	43054-45-1
50-32-8	62-54-4	82-28-0	97-74-5	110-44-1	134-72-5	477-30-5	636-23-7	1955-45-9	5800-19-1	17924-92-4	51325-35-0
50-33-9	62-55-5	82-68-8	97-77-8	110-57-6	135-20-6	488-41-5	636-79-3	2104-09-8	5834-17-3	18413-14-4	51410-44-7
50-44-2	62-56-6	83-59-0	98-01-1	110-85-0	135-23-9	493-78-7	637-07-0	2113-61-3	5979-28-2	18523-69-8	51542-33-7
50-55-5	62-73-7	83-79-4	98-85-1	110-89-4	135-88-6	504-88-1	671-16-9	2122-86-3	5989-27-5	18559-94-9	51630-58-1
50-78-2	62-75-9	84-65-1	98-92-0	111-44-4	136-40-3	509-14-8	683-50-1	2163-79-3	6109-97-3	18662-53-8	51786-53-9
50-81-7	63-25-2	85-44-9	98-96-4	111-46-6	137-17-7	510-15-6	712-68-5	2164-09-2	6119-92-2	18883-66-4	52207-83-7
51-03-6	63-92-3	85-68-7	99-30-9	112-27-6	137-26-8	512-56-1	720-69-4	2185-92-4	6120-10-1	18968-99-5	52214-84-3
51-21-8	64-17-5	86-06-2	99-55-8	113-92-8	139-05-9	513-37-1	756-79-6	2227-13-6	6151-25-3	19767-45-4	53609-64-6
51-28-5	64-75-5	86-29-3	99-56-9	114-83-0	139-13-9	517-28-2	758-17-8	2243-62-1	6294-89-9	19834-02-7	53757-28-1
51-55-8	64-77-7	86-30-6	99-57-0	114-86-3	139-40-2	518-75-2	759-73-9	2302-84-3	6334-11-8	20265-96-7	54143-56-5
51-75-2	66-05-7	86-50-0	99-59-2	115-02-6	139-61-1	520-18-3	760-60-1	2303-16-4	6358-85-6	20325-40-0	54150-69-5
51-79-6	66-27-3	86-57-7	100-00-5	115-07-1	139-94-6	520-45-6	765-34-4	2318-18-5	6369-59-1	20570-96-1	54749-90-5
52-24-4	67-20-9	86-74-8	100-40-3	115-28-6	140-11-4	525-66-6	772-43-0	2425-06-1	6373-74-6	20917-49-1	55090-44-3
53-19-0	67-21-0	86-86-2	100-41-4	115-29-7	140-49-8	531-06-6	785-30-8	2432-99-7	6381-77-7	21308-79-2	55268-74-1
53-70-3	67-48-1	86-87-3	100-42-5	115-32-2	140-56-7	531-18-0	828-00-2	2438-88-2	6385-58-6	21340-68-1	55556-92-8
53-95-2	67-52-7	86-88-4	100-44-7	115-96-8	140-57-8	531-82-8	834-28-6	2439-10-3	6452-73-9	21416-87-5	55557-00-1
53-96-3	67-66-3	87-29-6	100-51-6	116-06-3	140-67-0	531-85-1	838-88-0	2465-27-2	6959-47-3	21436-96-4	55567-81-2
54-11-5	67-72-1	87-51-4	100-52-7	116-29-0	140-79-4	532-32-1	842-00-2	2475-45-8	6959-48-4	21436-97-5	55738-54-0
54-12-6	67-98-1	87-56-9	100-63-0	117-10-2	140-88-5	536-33-4	842-07-9	2489-77-2	6965-71-5	21498-08-8	56222-35-6
54-31-9	68-23-5	87-68-3	100-75-4	117-39-5	141-90-2	538-41-0	860-22-0	2578-75-8	7008-42-6	21638-36-8	56654-52-5
54-80-8	68-76-8	87-86-5	100-97-0	117-79-3	142-04-1	540-23-8	868-85-9	2611-82-7	7227-91-0	21884-44-6	56795-65-4
54-85-3	68-89-3	88-19-7	101-05-3	117-80-6	142-47-2	541-69-5	869-01-2	2629-59-6	7347-49-1	22571-95-5	56795-66-5
55-18-5	69-65-8	88-73-3	101-14-4	117-81-7	142-59-6	542-75-6	915-67-3	2698-41-1	7411-49-6	22760-18-5	56894-91-8
55-22-1	70-25-7	88-85-7	101-21-3	118-74-1	143-19-1	542-88-1	924-16-3	2757-90-6	7422-80-2	22839-47-0	57497-29-7
55-31-2	71-43-2	88-96-0	101-54-2	118-75-2	143-50-0	548-62-9	924-42-5	2783-94-0	7519-36-0	23031-25-6	57497-34-4
55-80-1	71-55-6	89-25-8	101-61-1	118-92-3	147-24-0	551-92-8	930-55-2	2784-94-3	7572-29-4	23135-22-0	57653-85-7
55-98-1	72-20-8	90-43-7	101-73-5	119-34-6	148-18-5	553-53-7	932-83-2	2832-40-8	7631-99-4	23255-69-8	60102-37-6
56-04-2	72-33-3	90-94-8	101-79-1	119-38-0	148-24-3	555-84-0	937-25-7	2835-39-4	7632-00-0	23950-58-5	60391-92-6
56-23-5	72-43-5	91-53-2	101-80-4	119-53-9	148-79-8	556-52-5	938-73-8	2921-88-2	7681-93-8	24382-04-5	60599-38-4
56-38-2	72-54-8	91-59-8	101-90-6	120-36-5	148-82-3	563-41-7	943-41-9	3012-65-5	7757-82-6	24554-26-5	61034-40-0
56-49-5	72-55-9	91-62-3	102-09-0	120-61-6	149-29-1	563-47-3	952-23-8	3031-51-4	7758-19-2	25081-31-6	61702-44-1
56-53-1	72-56-0	91-76-9	102-50-1	120-62-7	149-30-4	569-57-3	959-24-0	3068-88-0	8065-91-6	25168-26-7	63412-06-6
56-72-4	73-22-3	91-79-2	102-71-6	120-71-8	150-38-9	569-61-9	961-11-5	3096-50-2	10024-97-2	25843-45-2	63885-23-4
56-75-7	74-31-7	91-93-0	102-77-2	120-78-5	150-68-5	576-68-1	968-81-0	3148-73-0	10048-13-2	26049-68-3	63886-77-1
56-86-0	74-96-4	91-94-1	103-03-7	120-80-9	151-56-4	578-76-7	971-15-3	3165-93-3	10102-43-9	26049-69-4	64049-29-2
57-06-7	75-00-3	92-13-7	103-16-2	120-83-2	156-10-5	590-21-6	999-81-5	3276-41-3	10318-26-0	26049-70-7	65734-38-5
57-14-7	75-01-4	92-52-4	103-23-1	120-93-4	156-51-4	592-31-4	1011-73-0	3296-90-0	10473-70-8	26049-71-8	67730-10-3
57-39-6	75-07-0	92-55-7	103-33-3	121-14-2	156-62-7	593-60-2	1068-57-1	3458-22-8	10589-74-9	26541-51-5	67730-11-4
57-41-0	75-09-2	92-67-1	103-72-0	121-66-4	262-12-4	593-70-4	1072-53-3	3544-23-8	12663-46-6	26628-22-8	68107-26-6
57-43-2	75-21-8	92-69-3	103-90-2	121-69-7	271-89-6	597-25-1	1078-38-2	3546-10-9	12789-03-6	28314-03-6	69658-91-9
57-50-1	75-25-2	92-84-2	104-46-1	121-75-5	297-76-7	598-55-0	1114-71-2	3564-09-8	13010-07-6	28322-02-3	72254-58-1
57-55-6	75-27-4	92-87-5	105-11-3	121-88-0	297-78-9	598-64-1	1116-54-7	3567-69-9	13010-08-7	28754-68-9	73785-40-7
57-57-8	75-34-3	93-46-9	105-36-2	122-34-9	298-00-0	602-87-9	1119-68-2	3570-75-0	13010-10-1	29082-74-4	74920-78-8
57-74-9	75-35-4	93-72-1	105-55-5	122-42-9	298-18-0	607-35-2	1120-71-4	3688-53-7	13073-35-3	30310-80-6	75198-31-1
57-97-6	75-56-9	93-76-5	105-60-2	122-60-1	302-15-8	608-73-1	1133-64-8	3693-22-9	13171-21-6	30418-53-2	75411-83-5
58-08-2	75-88-7	94-11-1	105-85-5	122-66-7	302-22-7	609-20-1	1146-71-0	3761-53-3	13256-11-6	31873-81-1	75881-18-4
58-14-0	76-01-7	94-20-2	106-46-7	123-31-9	302-79-4	611-23-4	1150-37-4	3775-55-1	13366-73-9	32221-81-1	75881-20-8
58-89-9	76-44-8	94-26-8	106-47-8	123-33-1	303-34-4	611-32-5	1150-42-1	3778-73-2	13483-18-6	32607-00-4	75881-22-0
59-02-9	77-06-5	94-52-0	106-50-3	123-73-9	303-47-9	612-82-8	1156-19-0	3817-11-6	13552-44-8	32852-21-4	75896-33-2
59-05-2	77-65-6	94-58-6	106-87-6	123-91-1	305-03-3	613-50-3	1162-65-8	3851-16-9	13743-07-2	33229-34-4	76180-96-6
59-33-6	77-79-2	94-59-7	106-88-7	124-48-1	306-37-6	613-94-5	1163-19-5	3883-43-0	13752-51-7	33857-26-0	77337-54-3
59-35-8	77-83-8	94-75-7	106-89-8	124-64-1	309-00-2	614-00-6	1212-29-9	4075-79-0	13838-16-9	33868-17-6	82018-90-4
59-51-8	78-34-2	94-80-4	106-92-3	126-72-7	315-18-4	614-95-9	1241-27-6	4106-66-5	14026-03-0	34176-52-8	86451-37-8
59-67-6	78-42-2	95-06-7	106-93-4	126-85-2	315-22-0	615-28-1	1248-18-6	4164-28-7	15356-70-4	34522-69-5	88208-16-6
59-87-0	78-59-1	95-14-7	106-99-0	127-06-0	319-84-6	617-84-5	1453-82-3	4247-02-3	15481-70-6	34627-78-6	89911-78-4
59-88-1	78-87-5	95-33-0	107-06-2	127-18-4	320-67-2	619-17-0	1465-25-4	4342-03-4	15721-02-5	35449-36-6	89911-79-5
60-11-7	79-00-5	95-50-1	107-07-3	127-47-9	324-93-6	619-67-0	1508-45-8	4363-03-5	15879-93-3	36133-88-7	91308-69-9
60-13-9	79-01-6	95-74-9	107-13-1	127-69-5	330-54-1	621-64-7	1582-09-8	4463-22-3	15973-99-6	36702-44-0	91308-70-2
60-34-4	79-06-1	95-79-4	107-20-0	128-37-0	333-41-5	622-51-5	1596-84-5	4548-53-2	16071-86-6	37087-94-8	91308-71-3
60-35-5	79-11-8	95-80-7	107-30-2	128-44-9	363-17-7	624-18-0	1634-78-2	4553-89-3	16219-99-1	38434-77-4	92177-49-6
60-51-5	79-19-6	95-83-0	108-03-2	128-66-5	389-08-2	624-84-0	1701-77-5	4680-78-8	16301-26-1	38514-71-5	92177-50-9
60-56-0	79-34-5	96-09-3	108-05-4	129-15-7	398-32-3	625-89-8	1746-01-6	4812-22-0	16338-97-9	38571-73-2	
60-57-1	79-44-7	96-12-8	108-30-5	131-01-1	434-13-9	628-02-4	1777-84-0	4998-76-9	16568-02-8	39156-41-7	

Appendix B. Fragments statistically associated with carcinogenicity or lack thereof ($p<0.01$)

Fragment structure[a]	Concordant observations	Discordant observations	Activity
C.**2C~C~-C~*C.*C~	12	1	+
C.**2C~C~-C~*C*C*C*C~	12	1	+
C**2C~C~-C**5C~.C*C*C~	12	1	+
C**2-3C~.C~C**5C~.C~	12	1	+
C**2-3C~.C~C~*C*C*C~	12	1	+
C..--2C.C.	9	2	-
C**2C~C-*4S::.C*C~	9	2	-
C~-*2S::.C*C*C.*C~	9	2	-
C--2C.N--4N:C-C.	11	0	+
O--2C~~C.=O	9	2	-
O--2C:.C~*C~	11	2	-
C~-*2C,C*C.*C~.	12	1	+
N--2-3N:C.C.=O	13	0	+
N--2-3N:C.C-C..	9	0	+
C--2C.C.-C..-C	9	2	-
C.--2C.C.-C..-C	8	1	-
C--2C.C.-C.-C.	9	2	-
C.--2C..C..-C.	10	2	-
C--2-3C..C.C.-C.	10	2	-
C--2C.C..-C..	9	2	-
C--2C...C-C..	10	2	-
C-*2*3N.S~N~	12	1	+
S**2C~C.*N~	12	1	+
N**2C~.C.*S~	12	1	+
C~-*2C:O~	9	0	+
C--2N..C:	11	0	+
C.--2O.C:.	10	1	-
C~-*2N::C*C*C~-C~*C~	11	0	+
N**2C~.C~-C~*C*C*C~.	13	0	+
C**2S~C~-C~*C*C*C~.	11	0	+
C**2-3N~C~C~*C*C~	12	1	+
O**2C~.C~-C~*C~	11	0	+
O**2C~.C~-C~*N*C~.	10	0	+
C--2-3O,C..C.	11	2	-
C~-*2N,C*C*C~-C~~	12	1	+
C~-*2C.C.*C~.	8	1	-
O'-C:.	9	1	-
C.--2C.C.-C..-C..-C.	6	0	-
C--2-3C..C..C-C.	8	0	-
C--2C:.C-C..	10	1	-
O--2C:.C~*C~.	9	2	-
C~-*2Cl,C*C-*5Cl,C.*C~	10	2	-
N--2C..C.	7	0	-
C--2O.C.=O	6	0	-
O--2C~~C-C:.	6	0	-
C~-*2O.C~-Cl	9	2	-
C.--2O,C:.	6	0	-
N--2C:.C~*C*C*C~	7	0	-
C:--2O,C~*C*C~	6	0	-
C.--2C~~C:.	7	0	-

In this list, fragments that are similar but not identical or imbedded one in the other have not been eliminated. In few cases they could come from similar groups of molecules.

[a]Fragment coding: the fragment code is composed by a list of atoms and bonds. In particular, each atom is followed by the list of appended bonds and bonds with atoms following in the code. The first elements of such a list are the appended bonds, i.e., the bonds that link the atom with another one not in the fragment; the other elements are the bonds with atoms belonging to the fragment.

The appended bonds are indicated by the following symbols:

(.) single bond;
(:) double bond;
(~) aromatic and heteroaromatic bonds;
(') ionic bond.

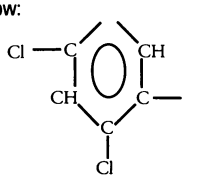
For example, C.: is equal to -C=.

The remaining bonds, with atoms belonging to the fragment, are indicated by the following symbols:

(-) single bonds;
(=) double bonds;
(*) aromatic and heteroaromatic bonds.

We have not described other bonds present in chemicals not appearing in Appendix B. To simplify the bonding information, we have used a unique symbol to codify the bonds involved in the main cyclic resonant structures (e.g., benzene, furan, thiophene). Each of the bonds of the second group may be followed by a number indicating the position in the fragment code of the second atom involved; if no number follows, then it is assumed that the bond is relevant to the immediately following atom. Note that the first atom in a fragment code is associated with the number zero, the second with the number 1, and so on. For example, in C--2N..C, the first C is linked to the N by a single bond and to the last C with another single bond. N has two appended single bonds.

Finally, it must be noted that the bonds with hydrogen are not indicated (they can be inferred from the valence of the atom), and a comma is simply used to separate two atoms not separated by other symbols. For example, C~-*2Cl,C*C-*5Cl,C.*C~ corresponds to the fragment below:



References

1. Klopman G, Rosenkranz HS. Testing by artificial intelligence: computational alternatives to the determination of mutagenicity. Mutat Res 272:59–71(1992).
2. Ashby J, Paton D. The influence of chemicals structure on the extent and sites of carcinogenesis for 522 rodent carcinogens and 55 different human carcinogen exposures. Mutat Res 286:3–74(1993).
3. Klopman G, Mc Gonigal M. Computer simulation of physical-chemical properties of organic molecules 1. Molecular system identification. J Chem Inf Comput Sci 21:48–52(1981).
4. Rosenkranz HS, Klopman G, Chankong V, Pet-Edwards J, Haimes YY. Prediction of environmental carcinogens: a strategy for the Mid-1980s. Environ Mutagen 6:231–258(1984).
5. Rosenkranz HS, Frierson MR, Klopman G. Use of structure-activity relationships in predicting carcinogenesis. In: Ong-term and short-term assays for carcinogens: a critical appraisal (Montesano R, Bartsch H, Vainio H, Wilbourn J, Yanasaki H, eds), IARC Scientific Publications No. 83. Lyon:International Agency for Research on Cancer, 1986; 497–517.
6. Klopman G, Frierson MR, Rosenkranz HS. The structural basis of the mutagenicity of chemicals in *Salmonella typhimurium*: the Gene-Tox Data Base. Mutat Res 228:1–50 (1990).
7. Rosenkranz HS, Klopman G. The structural basis of the mutagenicity of chemicals in *Salmonella typhimurium*: the National Toxicology Program Data Base. Mutat Res 228: 51–80(1990).
8. Rosenkranz HS, Klopman G. Structural basis of carcinogenicity in rodents of genotoxicants and non-genotoxicants. Mutat Res 228:105–124(1990).
9. Rosenkranz HS, Klopman G. New structural concepts for predicting carcinogenicity in rodents: an artificial intelligence approach. Teratog Carcinog Mutagen 10:73–88(1990).
10. Rosenkranz HS, Takihi N, Klopman G. Structure activity-based predictive toxicology: an efficient and economical method for generating non-congeneric data bases. Mutagenesis 6: 391–394(1991).
11. Klopman G, Rosenkranz HS. Structure–activity relations: maximizing the usefulness of mutagenicity and carcinogenicity databases. Environ Health Perspect 96:67–75(1991).
12. Gold LS, Sawyer CB, Magaw R, Backman GM, deVeciana M, Levinson R, Hooper NK, Havender WR, Bernstein L, Peto R, Pike MC, Ames BN. A carcinogenic potency database of the standardized results of animal bioassays. Environ Health Perspect 58:9–319(1984).
13. Gold LS, deVeciana M, Backman GM, Magaw R, Lopipero P, Smith M, Blumenthal M, Levinson R, Bernstein L, Ames BN. Chronological supplement to the carcinogenic potency database: standardized results of animal bioassays published through December 1982. Environ Health Perspect 67:161–200(1986).
14. Gold LS, Slone TH, Backman GM, Magaw R, Da Costa M, Lopipero P, Blumenthal M, Ames BN. Second chronological supplement to the carcinogenic potency database: standardized results of animal bioassays published through December 1984 and by the National Toxicology Program through May 1986. Environ Health Perspect 74:237–329(1987).
15. Gold LS, Slone TH, Backman GM, Eisenberg S, Da Costa M, Wong M, Manley NB, Rohrbach L, Ames BN. Third chronological supplement to the carcinogenic potency database: standardized results of animal bioassays published through December 1986 and by the National Toxicology Program through June 1987. Environ Health Perspect 84:215–286 (1990).
16. Ashby J, Tennant RW. Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NTP. Mutat Res 204:17–115(1988).
17. Ashby J, Tennant RW, Zeiger E, Stasiewicz S. Classification according to chemical structure, mutagenicity to Salmonella and level of carcinogenicity of a further 42 chemicals tested for carcinogenicity by the U.S. NTP. Mutat Res 223:73–103(1989).
18. Ashby J, Tennant RW. Classification according to chemical structure, mutagenicity to Salmonella and level of carcinogenicity of a further 39 chemicals tested for carcinogenicity by the U.S. NTP. Mutat Res 257:209–227 (1991).
19. Christofides N. Graph theory. London: Academic Press, 1975.
20. Balaban AT. Application of graph theory in chemistry. J Chem Inf Comput Sci 25:334–343(1985).
21. Cramer RD III, Redl G, Berkoff CE. Substructural analysis. A novel approach to the problem of drug design. J Med Chem 17: 533–535(1974).
22. Chu K.C, Feldmann RJ, Shapiro MB, Hazard Jr GF, Geran RI. Pattern recognition and structure-activity relationships studies. Computer-assisted prediction of antitumor activity in structurally diverse drugs in an experimental mouse brain tumor system. J Med Chem 18:539–545(1975).
23. Hodes L, Hazard GF, Geran RI, Richman S. A statistical-heuristic method for automated selection of drugs for screening. J Med Chem 20:469–475(1977).
24. Ashby J. Structural analysis as a means of predicting carcinogenic potential. Br J Cancer 37:904–923(1978)
25. Ashby J. Fundamental structural alerts to potential carcinogenicity or non carcinogenicity. Environ Mutagen 7:919–921(1985).
26. Klopman G, Kolossvary I. Evaluation of quantitative structure-activity predictions. Comparison of the predictive power of an artificial intelligence system with human experts. J Math Chem 5:389–401(1990).
27. Richard AM, Woo Y. A CASE-SAR analysis of polycyclic aromatic hydrocarbon carcinogenicity. Mutat Res 242:285–303(1990).
28. Bahler D, Bristol DW. The induction of rules for predicting chemical carcinogenesis in rodents. In: Intelligent systems for molecular biology (Hunter J, Shavlik J, Searls D, eds). Menlo Park, California:AAAI/MIT Press, 1993.
29. Ashby J, Tennant RW. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. Mutat Res 257:229–306(1991).
30. Tennant RW, Margolin BH, Shelby MD, Zeiger E, Haseman JK, Spalding J, Caspary W, Resnick M, Stasiewicz S, Anderson B, Minor R. Prediction of chemical carcinogenicity in rodents from in vitro genetic toxicity assays. Science 236:933–941(1987).
31. Klopman G, Rosenkranz HS. Quantification of the predictivity of some short-term assays for carcinogenicity in rodents. Mutat Res 253: 237–240(1991).
32. Gold LS, Bernstein L, Magaw R, Slone TH. Interspecies extrapolation in carcinogenesis: prediction between rats and mice. Environ Health Perspect 81:211–219(1989).
33. Huff J, Haseman J. Long-term chemical carcinogenesis experiments for identifying potential human cancer hazards: collective database of the National Cancer Institute and National Toxicology Program (1976-1991). Environ Health Perspect 96:23–31(1991).
34. Ashby J. Consideration of CASE predictions of genotoxic carcinogenesis for omeprazole, methapyrilene and azathioprine. Mutat Res 272:1–7(1992).
35. Tennant RW. Stratification of rodent carcinogenicity bioassay results to reflect relative human hazard. Mutat Res 286:111–118 (1993).